## **CLAIMS**

5 1. A method for estimating the selectivity of queries in a relational database, comprising the steps of:

constructing a probabilistic relational model (PRM) from said database; and

performing online selectivity estimation for a particular query.

10

15

25

- 2. The method of Claim 1, wherein sale PRM is constructed automatically, based solely on a data and space allocated to said PRM.
- 3. The method of Claim 1, wherein said selectivity estimation step further comprises the step of:

said selectivity estimator receiving as inputs both said query and said PRM, and outputting an estimate for a result size of said query.

- 4. The method of Claim 1, wherein the same PRM is used to estimate the size of a query over any subset of attributes in said database; and wherein prior information about a query workload is not required.
  - 5. The method of Claim 1, wherein selectivity estimation is performed for select queries over a single table; and wherein a Bayesian network is used to approximate joint distribution over an entire set of attributes in said table.

10

15

20

- 6. The method of Claim 1, wherein selectivity estimation is performed for queries over multiple tables; and wherein one or more PRMs are used to accomplish both select and join selectivity estimation in a single framework.
- 7. The method of Claim 1, further comprising the step of:
  learning PRMs with link uncertainty with a heuristic search algorithm.
  - 8. The method of Claim 7, wherein said search algorithm comprises a greedy hill-climbing search, using random restarts to escape local maxima.

9. A method for learning probabalistic relational models (PRM) having attribute uncertainty, comprising the steps of:

providing a parameter estimation task by

inputting a relational schema that specifies a set of classes, having attributes associated with said classes and having relationships between objects in different classes;

providing a fully specified instance of said schema in the form of a training database; and

performing a structure learning task to extract an entire PRM solely from said training database

- 10. The method of Claim 9, said structure learning task comprising the step of specifying which structures are candidate hypotheses.
- 25 11. The method of Claim 10, said structure learning task comprising the step of evaluating different candidate hypotheses relative to input data.

10

15

20

- 12. The method of Claim 11, said structure learning task comprising the step of searching hypothesis space for a structure having a high score.
- 13. A method for learning probabalistic relational models having link uncertainty, comprising the steps of:

providing a mechanism for modeling link uncertainty; and said mechanism computing sufficient statistics that include existence attributes without adding all nonexistent entities into a database.

14. The method of claim 10, sald mechanism comprising:

let  $\mu$  be a particular instant ation of Pa(X.E);

to compute  $C_{X.E}[true,\mu]$ , use a standard database query to compute how many objects  $x \in O^{\sigma}(X)$  have Pa(x.E);

to compute  $C_{X.E}[false, L]$ , compute the number of *potential* entities without explicitly considering each  $(x_1, ..., x_k) \in O'(Y_1) \times \bullet \bullet \bullet O'(Y_k)$  by decomposing the computation as follows:

let  $\rho$  be a reference slot of X with Range[ $\rho$ ] = Y;

let Pa(X.E) be the subset of parents of X.E along slot  $\rho$ ; and

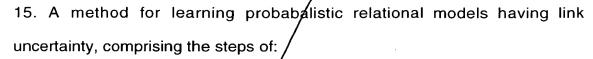
let  $\mu_{\rho}$  be a corresponding instantiation;

count a number of y consistent with  $\mu_{\rho}$ ;

if  $Pa_o(X.E)$  is/empty, this count is the |O'(Y)|;

wherein the product of these counts is the number of potential entities; to compute  $C_{X,E}[false,\mu]$ , subtract  $C_{X,E}[true,\mu]$  from said number.

25



providing a mechanism for modeling link uncertainty; and said mechanism computing sufficient statistics that include reference uncertainty, comprising the steps of:

fixing a set partition attributes  $\psi[\rho]$ ; and

treating a variable  $S_o$  as any other attribute in a PRM;

wherein scoring success in predicting a value of said attribute given a value of its parents is performed using standard Bayesian methods.

10

5